Compressing Massive Geophysical Data Sets Using Vector Quantization

Amy Braverman

California Institute of Technology

Overview:

- *This talk discusses a method for creating low-volume versions of massive geophysical data sets that approximately retain high-resolution data structure.*

Outline:

- Motivation.
- Strategy.
- Set-up and notation.
- Algorithm and its application.
- Example.

Strategy:

- **Partition** data using a spatio-temporal grid to create a family of data sets.
- **Summarize** each data set with a small set of records each of which contains a representative value (vector) and a count.
- Obtain the representatives and counts via a lossy data compression algorithm that optimally trades-off data corruption and data reduction.

15 of 36 MISR Channels

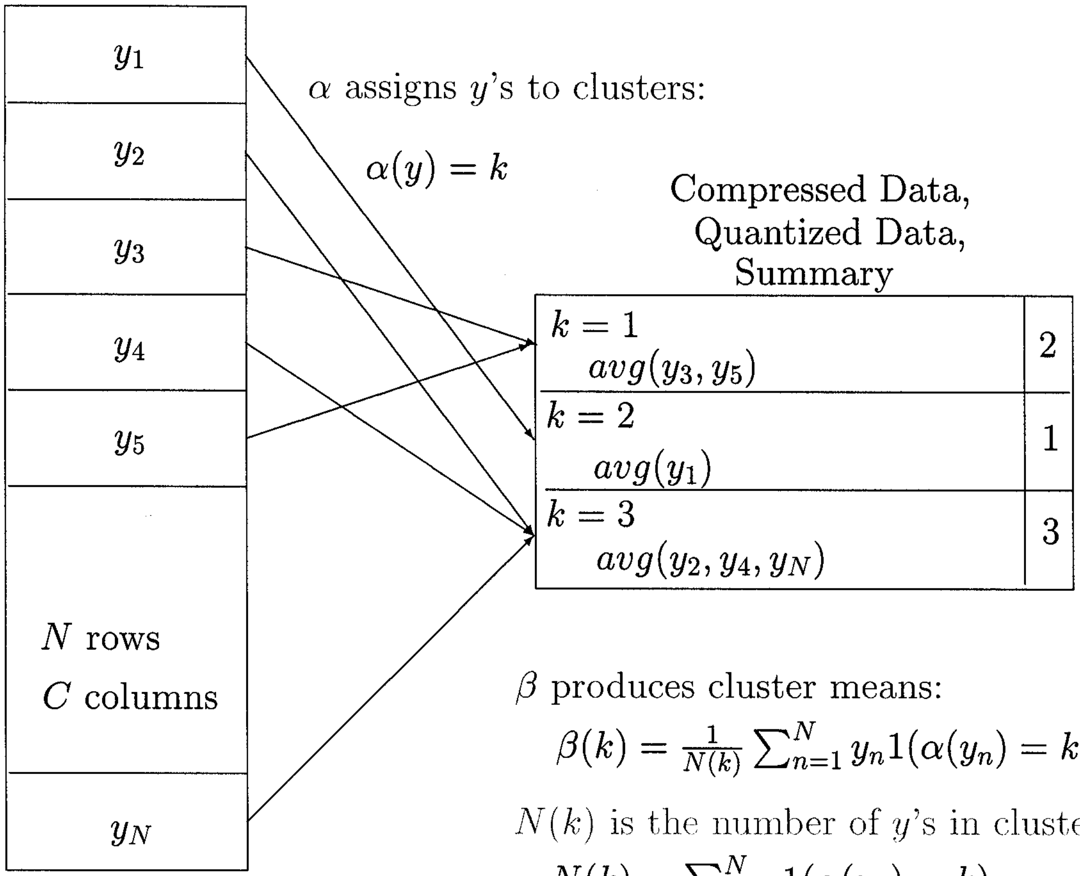
70.5° Forward.

45.0° Forward.

Nadir

45.0° Aft.

70.5° Aft.



β produces cluster means:

$$\beta(k) = \frac{1}{N(k)} \sum_{n=1}^{N} y_n 1(\alpha(y_n) = k)$$

$N(k)$ is the number of y's in cluster k:

$$N(k) = \sum_{n=1}^{N} 1(\alpha(y_n) = k)$$

The quantized value of y is the average of the cluster to which y belongs:

$$q(y) = \beta[\alpha(y)]$$

Two figures of merit for q, or equivalently, α:

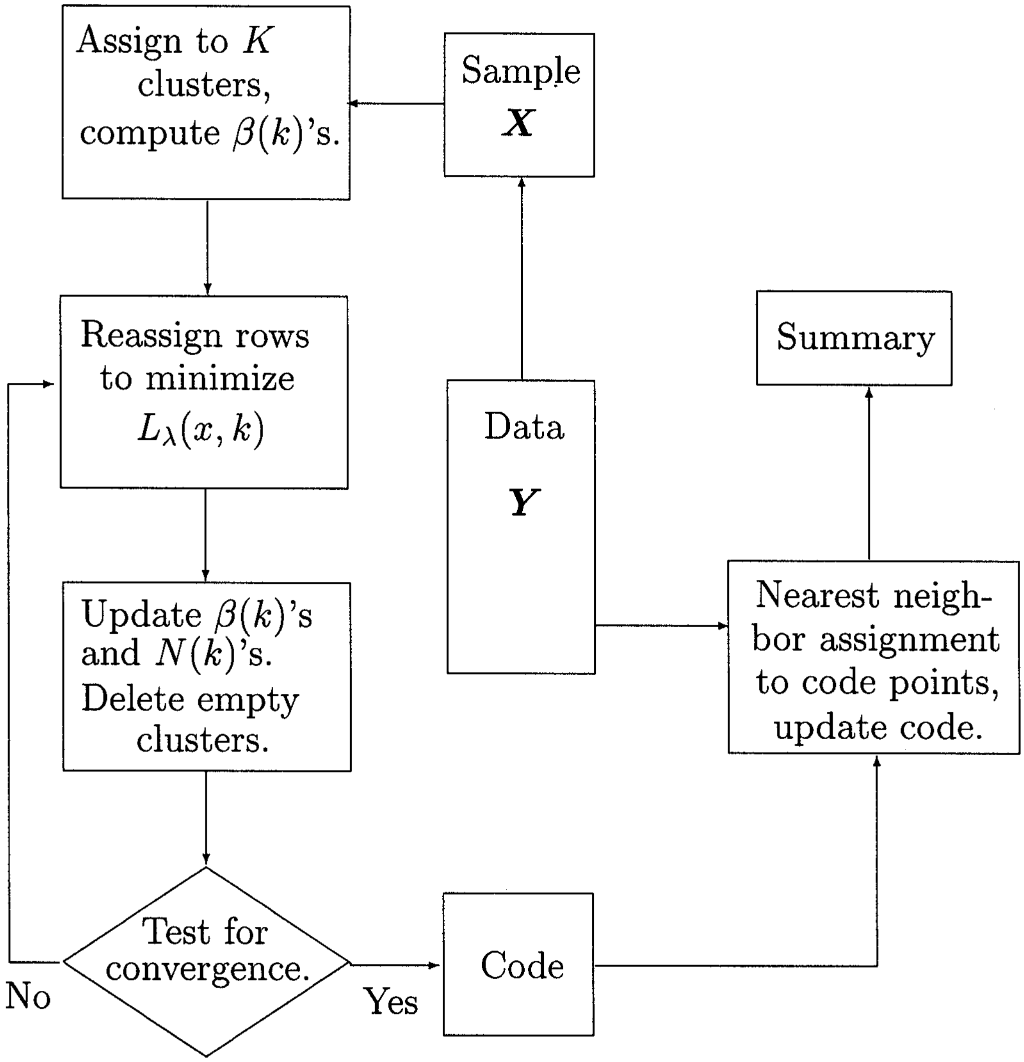
Distortion:

$$\delta(q) = \frac{1}{N} \sum_{n=1}^{N} \|y_n - q(y_n)\|^2$$

Entropy:

$$h(q) = -\sum_{k=1}^{K} \frac{N(k)}{N} \log \frac{N(k)}{N}$$

Extended ECVQ Algorithm



Assign to K clusters, compute $\beta(k)$'s.

Sample $\boldsymbol{X}$

Reassign rows to minimize $L_\lambda(x, k)$

Summary

Data $\boldsymbol{Y}$

Update $\beta(k)$'s and $N(k)$'s. Delete empty clusters.

Nearest neighbor assignment to code points, update code.

Test for convergence.

No

Yes

Code

$$L_\lambda(x, k) = \|x - \beta(k)\|^2 + \lambda \left[-\log \frac{N(k)}{N} \right]$$

Loss function: $\frac{1}{N} \sum_{n=1}^{N} L_\lambda(x_n, \alpha(x_n))$

Applying the algorithm:

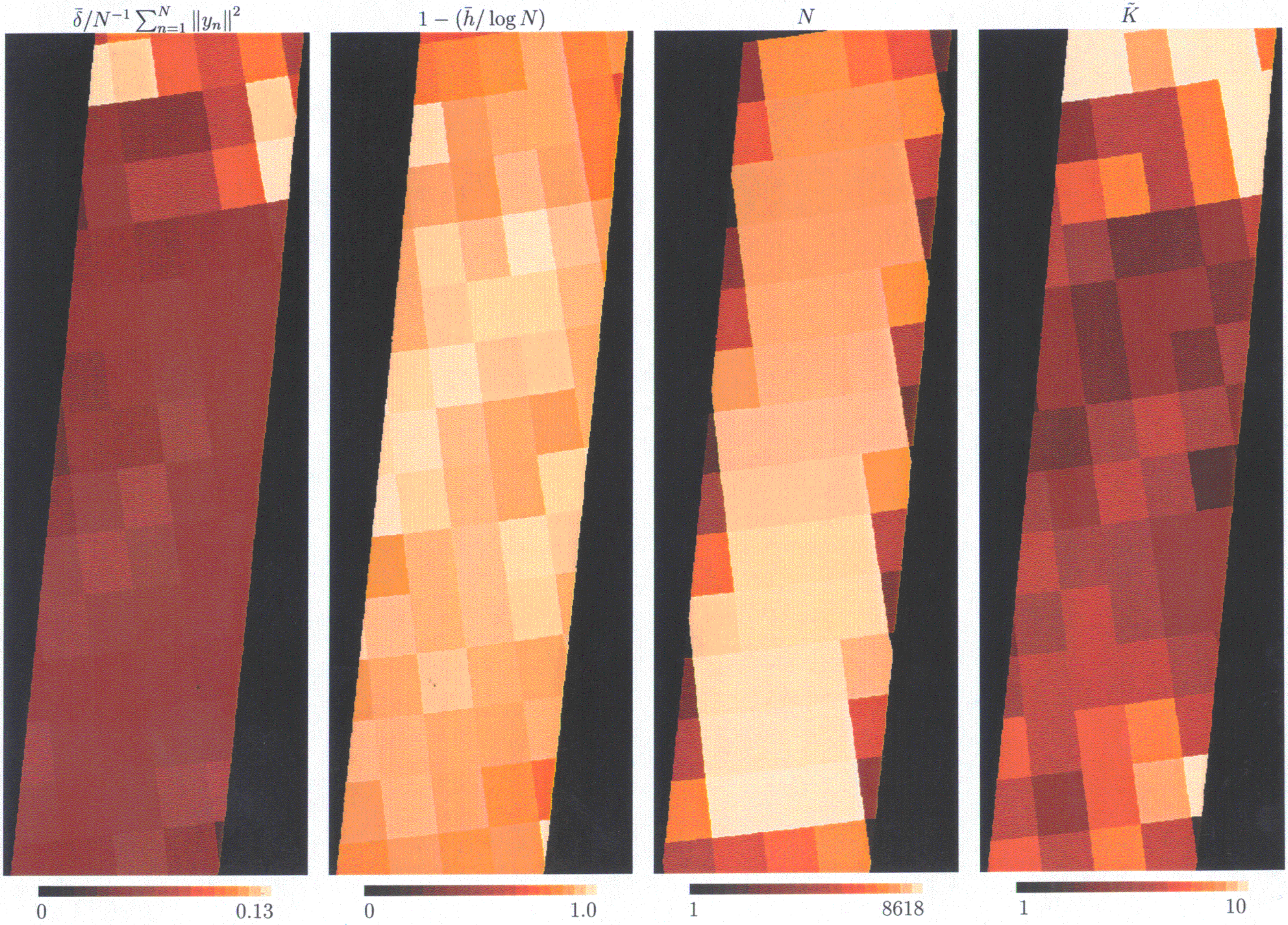
- ECVQ only summarizes the sample, not the full data set.
- ECVQ is subject to sampling variation.
- Solution: Use Extended ECVQ embedded in a Monte Carlo simulation (MCEECVQ) with S trials. Use different random samples on each trial. New figures of merit:

$$\bar{\delta} = \frac{1}{S}\sum_{s=1}^{S}\delta_s, \qquad \text{and} \qquad \bar{h} = \frac{1}{S}\sum_{s=1}^{S}h_s.$$

- In each cell, use the minimum distortion summary among the S generated to represent the cell.

Example:

- Apply MCEECVQ to 84 data sets separately using $K = 10$, samples of size 200, $S = 50$ trials.
- Report the summary with minimum δ.
- Result: summary data are 479 records (vs. 491,044); 36 variables.
- Processing time: about 8 minutes per cell on 195 MHz RISC 10000 processor.
- Analysis: K-means clustering to produce a thematic map.



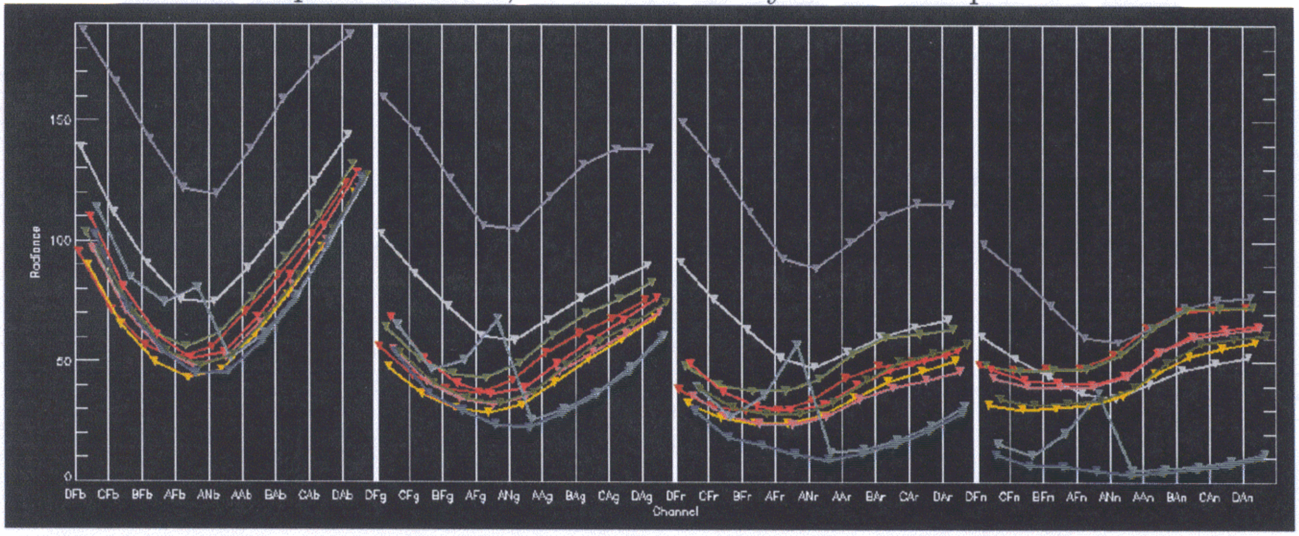
$\bar{\delta}/N^{-1}\sum_{n=1}^{N}\left\|y_n\right\|^2$

$1-(\bar{h}/\log N)$

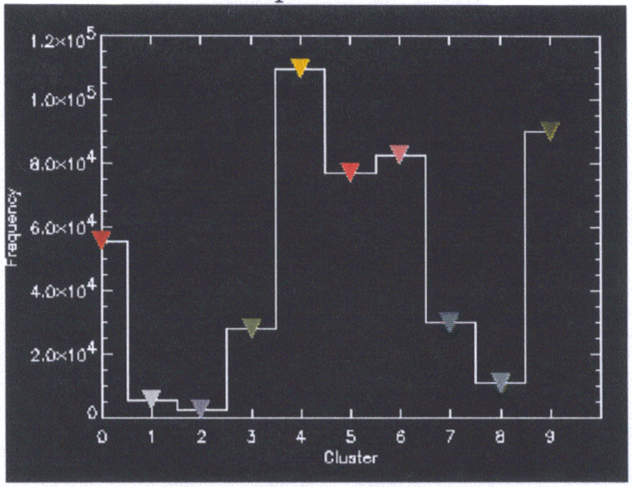
N

$\tilde{K}$

Cluster Representatives, *K*-means Analysis on Compressed Data



Cluster Populations, *K*-means Analysis
on Compressed Data


Frequency
2.0×10⁴
4.0×10⁴
6.0×10⁴
8.0×10⁴
1.0×10⁵
1.2×10⁵
Cluster